BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 379–391
Advance Access published on 19 June 2011
doi:10.1093/bib/bbr030

Computational methods for Gene Orthology inference

David M. Kristensen, Yuri I. Wolf, Arcady R. Mushegian and Eugene V. Koonin

Submitted: 28th March 2011; Received (in revised form): 5th May 2011

Abstract

Accurate inference of orthologous genes is a pre-requisite for most comparative genomics studies, and is also important for functional annotation of new genomes. Identification of orthologous gene sets typically involves phylogenetic tree analysis, heuristic algorithms based on sequence conservation, synteny analysis, or some combination of these approaches. The most direct tree-based methods typically rely on the comparison of an individual gene tree with a species tree. Once the two trees are accurately constructed, orthologs are straightforwardly identified by the definition of orthology as those homologs that are related by speciation, rather than gene duplication, at their most recent point of origin. Although ideal for the purpose of orthology identification in principle, phylogenetic trees are computationally expensive to construct for large numbers of genes and genomes, and they often contain errors, especially at large evolutionary distances. Moreover, in many organisms, in particular prokaryotes and viruses, evolution does not appear to have followed a simple 'tree-like' mode, which makes conventional tree reconciliation inapplicable. Other, heuristic methods identify probable orthologs as the closest homologous pairs or groups of genes in a set of organisms. These approaches are faster and easier to automate than tree-based methods, with efficient implementations provided by graph-theoretical algorithms enabling comparisons of thousands of genomes. Comparisons of these two approaches show that, despite conceptual differences, they produce similar sets of orthologs, especially at short evolutionary distances. Synteny also can aid in identification of orthologs. Often, tree-based, sequence similarity- and synteny-based approaches can be combined into flexible hybrid methods.

Keywords: *homolog; ortholog; paralog; xenolog; orthologous groups; tree reconciliation; comparative genomics*

INTRODUCTION

Identification of orthologous genes is a foundation of almost every comparative-genomic study. Orthologous gene sets are used to obtain information about evolutionary conservation and variability of molecular sequences, the tempo and mode of gene gain and loss, and constitute 'parts lists' for system-wide biological modeling. In comparative genomic studies, millions of genes in the now numerous sequenced genomes [1] cannot be considered completely independent of one another. Instead, sets of (putative) orthologous genes—in essence, instances of 'the same gene' in different species—are used to explore evolutionary histories and to utilize functional information about well-studied genes for annotation of their uncharacterized homologs [2–5].

Orthology, a term coined by Walter Fitch in 1970, refers to a specific type of relationship between homologous characters that arose by speciation at their most recent point of origin [6]. Here we restrict our focus to consider only genes, although the concept of orthology applies to other types of characters

Corresponding author. Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. E-mail: koonin@ncbi.nlm.nih.gov

David Kristensen is a Postdoctoral Fellow at the National Center for Biotechnology Information (NLM, NIH, Bethesda, Maryland, USA).

Yuri I. Wolf is an Associate Investigator at the National Center for Biotechnology Information (NLM, NIH, Bethesda, Maryland, USA).

Arcady R. Mushegian is Director of Bioinformatics Research at the Stowers Institute for Medical Research (Kansas City, Missouri, USA) and Professor of Microbiology at the University of Kansas Medical center (Kansas City, Kansas, USA).

Eugene V. Koonin is a Senior Investigator at the National Center for Biotechnology Information (NLM, NIH, Bethesda, Maryland, USA).

as well, such as chromosomal segments [7]. The problem of identification of orthologous genes is to distinguish between genes that are orthologous versus those that share another kind of homologous relationship such as paralogy [8]. The most common types of homologous relationships between genes are defined in Box 1. The events of the past, in particular speciation and gene duplication, cannot be observed directly but can be inferred, using algorithmic and statistical methods, from the genomic data available today. Thus, identification of orthology, even when highly confident, is technically always an inference.

Orthologs tend to retain similar molecular and biological functions [9]. In contrast, paralogs tend to diverge over time to perform different functions via subfunctionalization or neofunctionalization routes [10, 11]. However, functional conservation among orthologs should be inferred with caution because some orthologous genes can diverge functionally even among closely related organisms [12]. The reverse is also true: isofunctional genes are not necessarily related by orthology [13, 14].

Orthology has been originally defined for pairwise relationships between characters [6, 15], but in practice it is sets of orthologs from multiple species rather than individual orthologous pairs that are most often used to study the evolution of gene families and the organisms they reside in. Genes have different types of homologous relationships to different other genes—in a textbook example, human myoglobin is orthologous to mouse myoglobin, but paralogous to both mouse and human hemoglobins. More generally, as shown for the example in Figure 1, gene 1α in species C and gene 1 in A are orthologous because they are related by speciation at their point of origin in the last common ancestor at the base of the tree, and gene 1 in species A and gene 1β in C are similarly orthologous, whereas genes 1α and 1β in C are not orthologous, but rather paralogous as they are related at their most recent point of origin by a duplication event. Large-scale demarcation of orthologous and paralogous genes using pre-defined sets of probable orthologs is important for pinpointing key events in evolution and the associated shifts in molecular functions. For example, this approach has been employed to delineate the set of ancestral duplications in eukaryotes which showed significant excess of duplications among certain functional classes of genes [16].

Identification of genome-wide sets of orthologous and paralogous genes for distantly related organisms is a daunting task, because of the complexity of the routes of gene evolution that often involves horizontal gene transfer, lineage-specific gene loss, gene

Box I: Relationships between genes

- Homology: genes that share a common origin.
- Analogy: non-homologous genes that perform the same function as a result of convergent evolution.
- Orthology: genes arising by speciation at their most recent point of origin.
- Paralogy: genes arising by duplication at their most recent point of origin.
- Xenology: genes arising by HGT from another organism.
- In-/Out-paralogy: paralogous genes arising from lineage-specific duplication(s) after/before a given speciation event.
- Co-orthology: in-paralogous genes that are collectively, but not individually, orthologous to genes in other lineages (due to their common origin by speciation).
- Orthologous group: collection of all descendants of an ancestral gene that diverged from (after) a given speciation event.

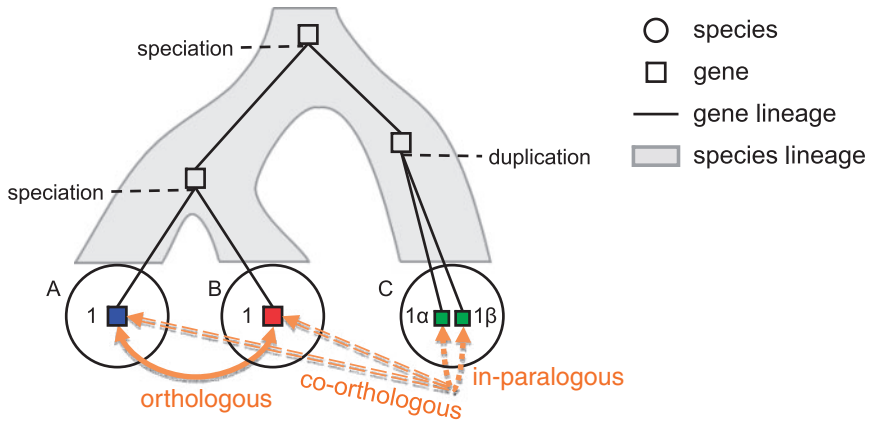

Figure I: Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.

fusion and fission, and other events that complicate evolutionary scenarios. At a time when the number of available complete genomes grows rapidly, it is also an important and increasingly urgent problem as reflected in the recent launch of the 'Quest for orthologs' initiative aiming at comparison and benchmarking of various existing methods for orthology detection [17]. In this review we touch only briefly on developing proper definitions of orthology, paralogy and other concepts and terms relevant to the evolutionary history of homologous genes, as well as applications of orthology detection methods, in order to concentrate on the computational approaches for detection of orthologous genes in genome sequences.

METHODS FOR IDENTIFICATION OF ORTHOLOGY

Phylogenetic tree-based approaches

Tree-based methods use an explicit model of the evolutionary history of the genes in question, in the form of a gene family tree, to infer orthologs. The most direct approaches compare this information with a second explicit evolutionary model of the organisms the genes reside in, i.e. a species tree, and use the procedure known as tree reconciliation or tree mapping [18, 19] to compare these two models to identify orthologs (Figure 2). The major assumption underlying this approach is that, by virtue of parsimony, the smallest number of evolutionary events (such as gene duplication or gene loss) is likely to reflect the actual course of evolution. Once the gene tree is constructed, orthologs and paralogs can be assigned by noting that paralogs group more closely together with members of the same species (Figure 2b), whereas orthologs group with members from other species (Figure 2c). A precise mapping function and an exact algorithm to sort duplication and speciation events have been developed [20]. This method formalizes the following intuition: if the offspring of a node in a gene tree is distributed among a given set of species, and the offspring of its direct descendant node is distributed among the same set of species (or a subset thereof), then there were no speciation events between the two nodes, so the former node is a duplication.

The tree construction step usually involves either distance-based (neighbor-joining and UPGMA) or character-based [maximum parsimony, various

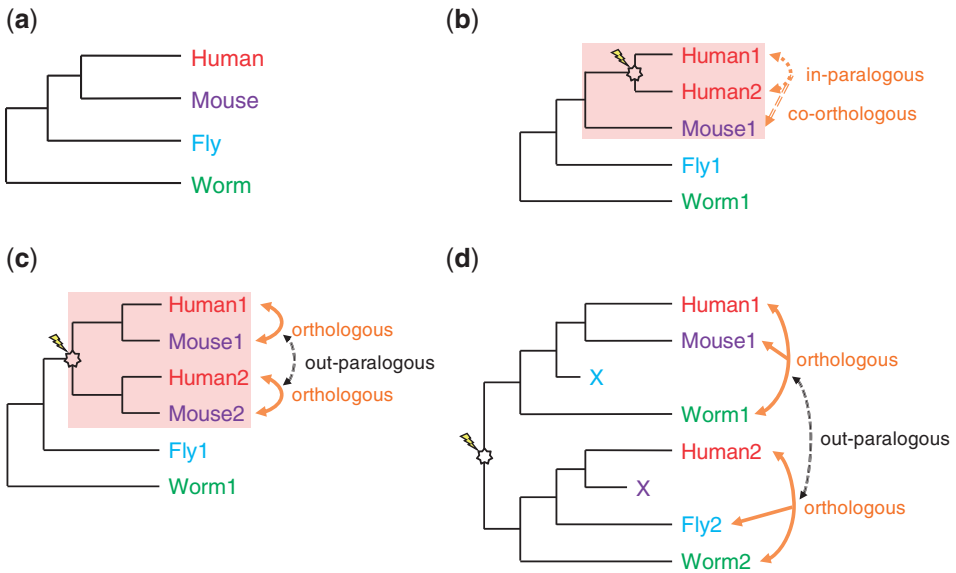

Figure 2: The reconciliation of the species tree (**a**) with an instance of a gene tree (**b–d**) allows for inference as to when evolutionary events such as speciation (T-branch), gene duplication (star-branch), or gene loss (X) occurred. (b) Gene tree with recent duplication, and evolutionary relationships shown for the genes in the shaded area. Because all three genes diverged from a single common ancestor, they would form a single orthologous group. (c) Gene tree with duplication preceding speciation event and evolutionary relationships shown for the genes in the shaded area. These four genes form two separate orthologous groups, corresponding to the two ancestral genes leading to each distinct gene lineage (Human1 and Mouse1, and Human2 and Mouse2). (d) Gene tree with duplication prior to speciation, followed by differential gene loss of Fly1 & Mouse2, where again all of the descendants of each of the two ancestral genes form an orthologous group.

kinds of maximum likelihood (ML), or Bayesian] algorithms. The distance matrix-based methods are much faster but limited in their applicability—in particular, they are less accurate when dealing with large distances or lineages with different rates of divergence [10]. Approximations of the ML approach are becoming available that help offset the otherwise high computational cost [21–24]. A major advantage of the tree-based approach to computational identification of orthologs is that it can use the information contained in a multiple sequence alignment, and can therefore model the evolution of the entire group of genes at once (in, for instance, a ML framework). Thus, the tree approaches are less prone to error than the pairwise heuristic approaches in situations such as differential gene loss [25, 26] (shown in Figure 2d).

In principle, explicit phylogenetic analysis is the most appropriate method for disentangling orthologous and paralogous genes, but there are several practical disadvantages to using trees. Trees are computationally expensive to produce when the number of leaves (organisms and genes) is large, and even though these can be produced and stored in large-scale databases with uses extending beyond orthology identification [27], any phylogenetic inference is also sensitive to noise and biases in the data [28, 29]. Probably the best-known artifacts are long- and short-branch attraction at large or small evolutionary distances, respectively [30]. Furthermore, tree construction is sensitive to the accuracy of multiple sequence alignment [23, 28, 31], which cannot be guaranteed when automated methods are used, especially when dealing with multi-domain proteins, a larger number of sequences and at larger evolutionary distances [32, 33]. Also, many tree construction methods treat as missing data columns in the alignment of the gene sequences that contain gaps. This approach reduces (in some cases drastically) the amount of information with which to create the model of evolution represented in the tree, and may introduce bias in this treatment of insertion and deletion events that have occurred during the evolution of a group of genes [32]. Even prior to constructing the multiple sequence alignment, the selection of homologs to align and build trees for must be performed. It is generally both impractical and undesirable to use all available sequences in a gene family for phylogenetic tree construction, not only because there are too many to apply the most reliable phylogenetic methods but also because different taxa are always unevenly represented. Any selection procedure has the potential to introduce biases which for large families may be substantial and exacerbate the technical problems of alignment and tree construction. Taken together, these difficulties preclude the application of phylogenetic analysis for the entire set of more than 1000 available complete genomes of diverse prokaryotes and eukaryotes [1].

A more fundamental challenge to the tree-based orthology analysis is presented by the fact that outside of multicellular eukaryotes, and especially in prokaryotes and viruses, evolution does not appear to have followed a 'tree-like' mode [34–38]. On the contrary, far from being a minor nuisance complicating the central trend of evolution, horizontal gene transfer (HGT) is a major component of the evolution of these organisms [39–45], so that their evolutionary history has to be represented by graphs that include not only vertical but also horizontal branches; algorithms for mapping of speciation and gene duplication events in such complex graphs are still unavailable.

A variety of computational platforms for orthology and paralogy detection and analysis have been developed to study the groups of organisms and gene families that have not been subject to substantial HGT, particularly those of animals, plants and fungi (these methods are also applied to prokaryotes and viruses, but the impact of HGT in these lineages on orthology assignment have not yet received sufficient attention). Some of the most advanced and widely used procedures for automated whole-genome phylogenomic identification of orthology are listed in Table 1, with the discussion of the methods that also use synteny information deferred until later. Many other methods exist, particularly those that rely on specialized databases of pre-computed orthologs for individual organisms or lineages, such as the Yeast Gene Order Browser [62], which further lists paralogs related by whole-genome duplication events ('ohnologs'). Many methods attempt to reduce the dependence on a single tree topology in various ways, whereas others abandon the strict reliance on trees entirely and instead use alternate, but similar, measures of sequence relatedness. Some methods do not attempt to provide a single prediction of orthologs or orthologous groups, but rather use multiple overlapping definitions of orthology. Probably the largest phylogenetic repository is PhylomeDB, which provides alignments, trees and

Table 1: Automated methods for phylogenomic prediction of orthology

Method	Description	Applied to
Orthostrapper/hierarchical grouping of orthologous and paralogous sequences (HOPS)	Uses bootstrap trees to calculate orthology support values for pairs of sequences in a multiple sequence alignment; graphical visualization by OrthoGUI [46].	Worms and mammals [47], and later to the domains of eukaryotes that appear in Pfam [48].
Resampled inference of orthologs (RIO)	Uses speciation duplication inference (SDI) algorithm [20] of fully-resolved bootstrap-resampled trees.	Pfam alignments of domains in plants and worms [49].
Réconciliateur d'Arbres Phylogénétiques (RAP)	Infers speciation and duplication events, and then identifies probable orthologs and paralogs in gene families with a given tree topology.	Databases of orthologous protein families HOVERGEN (dedicated to vertebrates), HOBACGEN (prokaryotes) and HOGENOM (organisms with completely sequenced genomes) [50].
COrrelation COefficient-based Clustering (COCO-CL)	Uses a measure of sequence distance between evolutionary histories of homologous genes instead of a species tree to construct a hierarchy of clusters.	Various protein classification databases (COGs [51], KOGs [52], OrthoMCL [53] and raw BLAST searches [54]).
Levels of Orthology From Trees (LOFT)	Constructs several hierarchical groupings that highlight different levels of relatedness between orthologs and paralogs.	Benchmarked against COGs, reconciliation with trusted species trees, and gene order conservation [55].
TreeFam	Uses several phylogenetic approaches to construct sets of orthologs from a curated resource of phylogenetic trees (extended with additional automatically generated trees).	Latest published release contains 25 fully sequenced genomes of animals plus four plant and fungal outgroup species [56].
Greenphyl	Uses semi-automatic gene family clustering to construct input dataset from raw data (all gene sequences in full genomes) prior to tree construction and an optimized stand-alone phylogenomic pipeline.	Complete plant genomes [57].
PhylomeDB	Uses a high-quality phylogenetic pipeline that includes evolutionary model testing and alignment trimming phases.	Latest published release contains 17 phylomes from such diverse organisms as human, yeast and even bacteria [58].
Berkeley PhyloFacts Orthology Group (PHOG)	Uses pre-computed trees, but allows targeting of different taxonomic distances and precision levels via user-set tree-distance thresholds in a prediction webserver.	Human, mouse, zebrafish and fruit fly sequences from TreeFam-A database [59].
MetaPhylogenyBased Orthologs (MetaPhOrs)	Applies a species overlap algorithm [60] to integrate information from multiple phylogenetic trees from a wide variety of sources.	Hundreds of genomes from eukaryotes and prokaryotes, although the authors note greatly reduced performance in the latter [61].

orthology predictions for every protein in each genome in its database (a 'phylome'), with its most recent release containing phylomes from 17 diverse organisms. MetaPhOrs, developed by the same authors, is probably the largest repository of phylogenetic orthology predictions.

Heuristic best-match methods

While phylogenetic tree analysis relies on an explicit model of the evolution of genes and species, an alternative class of approaches instead relies on the assumption that the sequences of orthologous genes (proteins) are more similar to each other than they are to any other genes from the compared organisms (Figure 3 and [63]; see below for the special case of in-paralogs). In practice, the use of symmetric best-match relationships [51] (often called BBHs, for bidirectional best hits [64]), is the most common method employed to infer probable orthologs in comparative genomic studies. The BBHs can be easily determined from the ranking of homologs obtained in a pairwise sequence similarity search, bypassing the need for reconciliation of phylogenetic trees. Many best-match algorithms go further, in a process sometimes called pair-linking, to group together genes from multiple genomes that are orthologous or co-orthologous to one another, so that these groups jointly represent all of the descendants of a common ancestral gene within the studied set of organisms [8]. Linking pairs of BBHs from multiple genomes has a property of self-verification, as their consistency would be highly unlikely due to chance, especially between phylogenetically distant lineages [51].

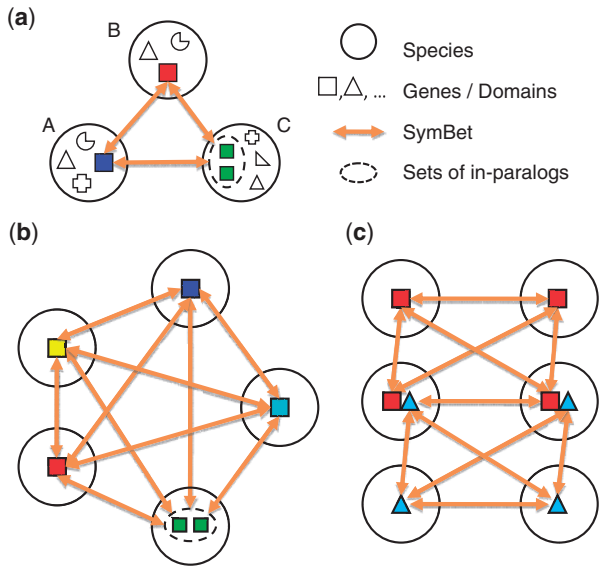

Figure 3: Grouping of genes in different species that are each others' BBHs into sets of orthologs and co-orthologs. (**a**) Graph representation of the evolutionary scenario shown in Figure 1, with genes represented as vertexes and BBHs as edges. (**b**) A larger, completely connected orthologous group of six genes from five species. (**c**) An even larger group that contains some members that are not orthologs, in this case due to domain recombination, where the top members have one domain and the bottom members have another, non-homologous domain, but they were merged into the same group due to the middle members containing both domains (thus bridging the otherwise disconnected components in the BBH graph). Alternative scenarios of improper merging can involve differential gene loss or large, complex mixtures of in- and out-paralogs, but in all three cases are due to the pair-wise procedure used to add members to groups, without considering the long-range relationships to the other members in the group.

Pairwise BBH relationships are usually determined by taking the top-ranking matches found by BLAST [65], for which highly efficient implementations are available (such as NCBI [66] and WU-BLAST [67]), or by other sequence similarity measures such as the similarity scores computed from Smith-Waterman alignments [68] or ML distance estimates from significant scoring pair-wise alignments (reciprocal smallest distance, RSD [69]). The methods for clustering these pairwise relationships into orthologous groups vary, with the most widely used approach involving deterministic single-linkage clustering procedures, where any two clusters sharing a common BBH are merged until convergence [51].

Heuristic algorithms present a number of advantages over tree-based approaches [8, 70]. They are typically much faster, easier to automate, and a number of efficient implementations have been developed that can handle very large numbers of genomes. Given that such algorithms do not rely on either species trees or gene trees, they avoid the artifacts associated with constructing and using phylogenies (see above). Moreover, because these heuristic methods rely on the ranking of sequence similarity scores rather than on multiple alignments, they also avoid many of the pitfalls inherent to multiple sequence alignments and choosing lists of homologs that adversely affect the accuracy of phylogenetic tree analysis [32, 33].

Heuristic approaches to orthology identification are vulnerable to their own types of errors. In particular, pairwise associations typically fail to detect differential gene loss [71, 72]—for example, in the scenario illustrated in Figure 2d, the BBH assumption is false because even though Mouse1 and Fly2 are each others' highest-ranking matches in that pair of genomes, this is due to the differential loss of Fly1 and Mouse2 in their respective lineages rather than to a genuine orthologous relationship. In a case like this, a tree-based approach can pinpoint the lineage-specific gene loss by noting that a genome from another lineage contains both paralogs. In addition, the method of constructing orthologous groups from pairwise BBHs may be overly inclusive and create mixed groups that do not accurately represent the evolutionary history of the collection as a whole, especially in large, complicated families where an explicit model of the evolution undergone by these genes can help to identify different relationships. An additional source of erroneous orthology assignment is domain recombination. Figure 3c illustrates an extreme scenario where two groups of orthologs that each contain a distinct conserved domain but that do not share any regions of homology with one another, nevertheless can be merged due to other genes that contain both domains. Because lineage-specific gene loss and variable multidomain architectures, even among supposedly orthologous proteins, are particularly common in eukaryotes [73–75], several solutions have been devised to deal with these issues (discussed below).

It has also been suggested that the BLAST score might not be a good indicator of the actual evolutionary relationship between a pair of homologs, as illustrated for select cases where the top BLAST

Table 2: Automated methods for heuristic best-match prediction of orthology

Method	Description	Applied to
Clusters of orthologous groups (COGs), variants and derivatives	Identifies three-way BBHs between orthologs or sets of co-orthologs in three different species, and these groups expanded (merging triangles whenever they share a common side) until saturation, followed by manual splitting of large groups improperly joined by multidomain proteins or complex mixtures of in- and out-paralogs [51]; later developments focused on expanding the resource [77, 52], adding automation [78, 79], and more efficient handling of large numbers of genomes [80].	Initially the first seven completely-sequenced genomes available [51], with subsequent updates [77], expansions (including the automated eggNOG [78, 79] currently containing 630 complete genomes), and several lineage-specific derivatives including: eukaryotic KOGs [52], COGs for individual phyla [81, 82], archaeal arCOGs [83], dsDNA phage POGs [84], large nucleo-cytoplasmic DNA virus NCVOGs [85] and *n*-way BBHs in herpes viruses [86].
OrthoMCL	Forms groups of orthologs and co-orthologs using a Markov clustering process involving iterative simulations of stochastic (randomized) flow on the edges of a BBH graph, with clusters of desired tightness identified depending on a given 'inflation' parameter determined by trial and error [53, 87].	The first fully automated heuristic algorithm applied across multiple eukaryotic taxa [53]; the current OrthoMCL-DB [88] version 4 contains 138 genomes of mostly eukaryotes, but also some bacteria and archaea.
InParanoid/multiparanoid	Detects BBHs between a pair of organisms and then applies additional statistical rules to add in-paralogs arising from duplication after speciation [89, 90]; multiparanoid later developed to combine pairwise predictions into multi-species groups [91].	Latest 7.0 release covers 99 eukaryotic species plus *Escherichia coli* outgroup [92].
OMA	Various improvements upon traditional BBH strategies such as RSD evolutionary distances and accounting for differential gene loss and gene fusion–fission events [93–95].	One of the largest projects of its kind with 1000 genomes of both prokaryotic and eukaryotic species.
RoundUp	Uses ML-based evolutionary distances (RSD [69]) in pair-wise comparisons of hundreds of genomes [96].	Current version includes >900 genomes of both prokaryotic and eukaryotic species.
Domain-based detection of orthologs (DODO)	Efficient BBH approach based on domain architectures [97].	Benchmarked against InParanoid's 100 genomes.
OrthoInspector	First creates groups of in-paralogs and then examines 1-to-1, 1-to-many, or many-to-many reciprocal matches between pairs of groups, with additional detection of contradicting information between the two groups (such as an incomplete proteome) [98].	Fifty-nine eukaryotic organisms with approximately complete proteomes.

match is not the same as the nearest tree-neighbor, i.e. not a bona fide ortholog [76]. However, these examples notwithstanding, the BLAST score ranking appears to be a good statistical predictor of orthology at the genome scale, especially when BBHs rather than one-way best matches are used, and even more so when the BBHs share consistency with additional genomes (ARM, unpublished observations). Indeed benchmarking of algorithms for ortholog definition suggests good agreement between phylogenetic tree-based and heuristic best-match approaches (see below).

At this time, all their limitations notwithstanding, heuristic best-match approaches have managed to produce extensive collections of (putative) orthologous groups covering large numbers of species. The first to succeed at this task was the clusters of orthologous groups (COGs) method [51] that employed the prototype BBH pair-linking algorithm involving three-way symmetric best matches, merged with single-linkage clustering. Subsequent expansions, variations and derivatives of this method, as well as several other popular heuristic methods, are described in Table 2, again with the discussion of those that also use synteny information deferred until later.

The latest heuristic algorithms are gradually overcoming many of their hindrances. For instance, the use of domains rather than full genes avoids the domain recombination problem by more precisely defining the region of a gene that is orthologous [83, 84, 97], an approach that could benefit tree-based methods as well because multidomain architecture also complicates the choice of homologs and construction of a multiple sequence alignment. However, more algorithmic development is required to avoid improperly merging groups due to complex mixtures of differently-related genes,

such as in cases of differential gene loss (Figure 2d) [25, 26].

Synteny

The conservation of local gene order (synteny) is a consequence of common ancestry that is most often observed among closely-related organisms. About half of all orthologous genes in human and fish belong to conserved synteny blocks [99]. In vertebrates, synteny appears to be (nearly) evolutionarily neutral with a few exceptions [100, 101], although rates of genomic rearrangement are highly variable in different lineages [102]. Homologs surrounded by the sets of orthologous genes in these organisms are thus very likely to be orthologous themselves [103]. However, at least in animals, the rate of loss of syntenic neighborhoods is roughly proportional to the rate of amino acid sequence divergence in orthologs, and synteny becomes undetectable when the average protein identity is <50% [104, 105]. Prokaryotes show a still higher rate of synteny loss [106–109], which may occur even at >90% identity [100, 101] except in a relatively small fraction of conserved neighborhoods where selection pressure appears to act to retain gene order [110, 111].

In itself, synteny is not a powerful approach for orthology identification because gene orders generally evolve much faster than gene repertoires or protein sequences. Nevertheless, at close evolutionary distances, synteny can be used to support the confidence in orthology predictions, and even help to distinguish between orthology that has been maintained vertically throughout a gene's evolutionary history and xenology, resulting from HGT [112–115]. Synteny information has been combined with a phylogenetic tree approach in OrthoParaMap [116] and PhyOP [117] to measure orthology between a pair of closely related species, and in SYNERGY [118] to use this information when available among a large group of species. Synteny has also been combined with a BBH pair-linking approach in the alignable tight genomic clusters (ATGCs) across groups of closely-related prokaryotic genomes [119], and in MSOAR (subsequently extended to MultiMSOAR), a high-throughput ortholog assignment system based on genome rearrangement that has been applied to mammals [120, 121].

Hybrid and other approaches

Phylogenetic and heuristic approaches can be combined with each other or with synteny information, to yield hybrid approaches that attempt to overcome the shortcomings of using either method alone. For example, hybrid approaches can offset the computational expense of a phylogenetic approach, or reduce the vulnerability of heuristic algorithms to evolutionary events such as differential gene loss. Ortholuge uses a phylogenetic approach to refine clusters made by a heuristic algorithm, noting cases where relative gene divergence is atypical between two compared species and an outgroup species and therefore suggests paralogy rather than orthology [122]. EnsemblCompara further integrates the tree-reconciliation and BBH pair-linking approaches by starting with gene trees made from the initial clusters produced by heuristic algorithms, and reconciling these with the species tree of vertebrates [27]. HomoloGene is another hybrid approach that uses pairwise gene comparisons but follows a guide tree to compare more closely related organisms first, and also adds gene neighborhood conservation [1]. Other approaches also exist that do not fall into any of the above categories, including a method that uses topological distance in a species tree (which it does not reconcile with a gene tree) as a factor in a linkage equation to find dense clusters in a multipartite graph (whose edges are not restricted to BBHs) [123] and a machine-learning predictor of orthology using a set of graph features that, in addition to sequence similarity and synteny, also includes gene co-expression and protein interaction networks [124].

COMPARISONS OF ORTHOLOGY DETECTION METHODS

Several direct comparisons of computational methods to identify pairs of orthologs and orthologous groups have provided insight into which approaches work best in various contexts [3, 70, 125–127]. As there is no widely accepted 'gold-standard' set of orthologs, one of the authors of the OrthoMCL method developed a statistical approach that compared several methods of ortholog identification against one another [127]. Using a Latent Class Analysis technique [128], the overlap between several sets of eukaryotic orthologous groups made by different programs was analyzed in terms of sensitivity and specificity. By these measures, no single method achieved optimal performance; each method reaches a different trade-off between the two criteria. For instance, many BBH pair-linking methods have been found to reach high sensitivity

at the cost of specificity (larger groups containing unique and by inference dubious predictions not found by other methods, in particular when the arbitrary sequence similarity cutoffs were relaxed). The tree-based methods displayed the opposite trend (larger number of smaller groups with the predictions also found by most of the other methods in the study). The heuristic algorithms InParanoid [92] and OrthoMCL [88] exhibited the most even balance between the two (medium-sized groups).

Without a standard of truth, it is difficult to ascertain whether a given ortholog prediction is a true positive or a false positive, or whether a missed ortholog prediction is a false negative or a true negative [128, 129]. In the comparison of the methods against one another, differences are expected because genes can have different types of homologous relationships to different other genes (Figure 1), and thus methods that choose different speciation events to define co-orthology will produce different results. For example, a speciation event chosen prior to a whole-genome duplication event [130, 131] in one lineage will result in each of the duplicated pairs being grouped together as in-paralogs/co-orthologs that share a single common ancestor, as shown in Figures 1 and 2b. By contrast, if the duplication occurs prior to the speciation as in Figure 2c, then each pair will be separated into distinct groups of orthologs, as there were two copies in the common ancestor of the descendant species. The choice of this speciation event is also related to the purpose for which the sets of orthologs are built: for instance, if the goal is to construct groups where all of the genes perform the same function [3], then a more recent speciation event is chosen. If, however, the goal is to study the evolution of all of the descendants of a distant ancestral gene, then the groups necessarily contain genes that are more divergent in sequence and function, and include more in-paralogs/co-orthologs. Therefore, among the methods that group together orthologs and co-orthologs, the within-group consistency was also examined with respect to several additional factors such as gene function and domain architecture, where again OrthoMCL has been reported to perform better than other methods. Another study that assessed the feasibility of using orthology identification to predict similar functions among homologous genes by using functional genomics data, such as gene expression and protein interactions, has found (as one might intuitively expect) that the less inclusive methods (that produce smaller groups) retained a higher degree of functional similarity within those groups [3].

More recently, a larger study has been undertaken where again several methods were compared, this time with respect to both phylogeny and function (including some benchmarks from literature), and both eukaryotes and prokaryotes were examined [126]. One of the main, perhaps surprising results was that the more sophisticated tree reconstruction and reconciliation approach of EnsemblCompara [27] was sometimes outperformed by pairwise comparison approaches. These findings have been corroborated by evaluation of the functional similarity of predicted orthologs [using Gene Ontology (GO) annotations [132], enzyme classification (EC) numbers [133], correlation of expression level in human and mouse [134], and gene neighborhood conservation]. Another notable finding was that even a generic BBH approach often outperformed protocols with more complex algorithms. This, in addition to the speed and simplicity considerations, may help explain why many researchers prefer to use simple ad-hoc implementations of BBH rather than more sophisticated methods.

CONCLUSIONS

Identification of orthologous genes is an essential task in comparative genomics that is complicated by non-uniform evolutionary rates, extensive gene duplication, gene loss and horizontal gene transfer. The methods for inferring pairs or groups of orthologs fall into two main classes, the tree-based and heuristic best-match methods; gene synteny can also be used to aid in ortholog identification. Benchmarking analyses show that tree-based and heuristic methods in practice often yield similar sets of predicted orthologs, with the differences mostly due to the choice of speciation event used to define the co-orthologs/in-paralogs. The tree-based methods tend to be more specific whereas the heuristic methods are often more sensitive. Tree-based methods are preferable in principle as they employ explicit models of evolution that allow the classification of orthologs, co-orthologs, in-paralogs and out-paralogs. However, these methods are computationally expensive, prone to artifacts of multiple sequence alignment and phylogenetic inference, and may not perform well in cases of horizontal gene transfer. For large data sets, particularly in

prokaryotes where evolution does not follow a simple tree-like pattern, the more efficient and inclusive similarity-based heuristic methods are important.

Key Point

- Orthologs are typically identified by phylogenetic analysis, a heuristic similarity-based approach, or a combination of the two, with synteny information helpful to improve predictions when it is available. The tree-based approaches are the clear choice for small sets of animal and plant species, whereas heuristic approaches are required for datasets of thousands of genomes and for gene families that have undergone horizontal transfer rather than tree-like vertical descent, as is common in prokaryotes.

FUNDING

Intramural Research Program of the National Library of Medicine at the US National Institutes of Health; Stowers Institute for Medical Research.

References

1. Sayers EW, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2011;**39**(Database issue):D38–51.
2. Park D, Singh R, Baym M, *et al*. IsoBase: a database of functionally related proteins across PPI networks. *Nucleic Acids Res* 2011;**39**(Database issue):D295–300.
3. Hulsen T, Huynen MA, de Vlieg J, *et al*. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 2006;**7**(4):R31.
4. Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 1998;**8**(3):163–7.
5. Sjolander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 2004;**20**(2):170–9.
6. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool* 1970;**19**(2):99–113.
7. Hachiya T, Osana Y, Popendorf K, *et al*. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics* 2009;**25**(7):853–60.
8. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005;**39**:309–38.
9. Peterson ME, Chen F, Saven JG, *et al*. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci* 2009;**18**(6):1306–15.
10. Lynch M, Katju V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* 2004;**20**(11):544–9.
11. Ohno S. *Evolution by Gene Duplication*. New York: Springer, 1970.
12. Diaz R, Vargas-Lagunas C, Villalobos MA, *et al*. argC Orthologs from Rhizobiales show diverse profiles of transcriptional efficiency and functionality in *Sinorhizobium meliloti*. *J Bacteriol* 2011;**193**(2):460–472.
13. Omelchenko MV, Galperin MY, Wolf YI, *et al*. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct* 2010;**5**:31.
14. Henschel A, Kim WK, Schroeder M. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics* 2006;**22**(5):550–5.
15. Fitch WM. Homology a personal view on some of the problems. *Trends Genet* 2000;**16**(5):227–31.
16. Makarova KS, Wolf YI, Mekhedov SL, *et al*. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res* 2005;**33**(14):4626–38.
17. Gabaldon T, Dessimoz C, Huxley-Jones J, *et al*. Joining forces in the quest for orthologs. *Genome Biol* 2009;**10**(9):403.
18. Mirkin B, Muchnik I, Smith TF. A biologically consistent model for comparing molecular phylogenies. *J Comput Biol* 1995;**2**(4):493–507.
19. Page RD, Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol* 1997;**7**(2):231–40.
20. Zmasek CM, Eddy SR. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 2001;**17**(9):821–8.
21. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**(3):e9490.
22. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**(7):1641–50.
23. Liu K, Raghavan S, Nelesen S, *et al*. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 2009;**324**(5934):1561–4.
24. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003;**52**(5):696–704.
25. Hughes AL, Friedman R. Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc Biol Sci* 2004;**271(Suppl 3)**:S107–9.
26. Gout JF, Duret L, Kahn D. Differential retention of metabolic genes following whole-genome duplication. *Mol Biol Evol* 2009;**26**(5):1067–72.
27. Vilella AJ, Severin J, Ureta-Vidal A, *et al*. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 2009;**19**(2):327–35.
28. Felsenstein J. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, 2004.
29. Hahn MW. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 2007;**8**(7):R141.
30. O'Connor T, Sundberg K, Carroll H, *et al*. Analysis of long branch extraction and long branch shortening. *BMC Genomics* 2010;**11(Suppl 2)**:S14.
31. Thompson JD, Linard B, Lecompte O, *et al*. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 2011;**6**(3):e18093.
32. Liu K, Linder CR, Warnow T. Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr* 2010;**2**:RRN1198.

33. Thorne JL, Kishino H. Freeing phylogenies from artifacts of alignment. *Mol Biol Evol* 1992;**9**(6):1148–62.
34. Schliep K, Lopez P, Lapointe FJ, *et al*. Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol Biol Evol* 2011;**28**(4):1393–405.
35. Olendzenski L, Gogarten JP. Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. *Ann N Y Acad Sci* 2009;**1178**:137–45.
36. Doolittle WF. The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Philos Trans R Soc Lond B Biol Sci* 2009;**364**(1527):2221–8.
37. Bapteste E, O'Malley MA, Beiko RG, *et al*. Prokaryotic evolution and the tree of life are two different things. *Biol Direct* 2009;**4**:34.
38. Koonin EV, Wolf YI. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biol Direct* 2009;**4**:33.
39. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 2011;**7**(1):e1001284.
40. Puigbo P, Wolf YI, Koonin EV. The tree and net components of prokaryote evolution. *Genome Biol Evol* 2010;**2**: 745–56.
41. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 2008;**105**(29): 10039–44.
42. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 2005; **3**(9):679–87.
43. Boucher Y, Douady CJ, Papke RT, *et al*. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* 2003;**37**:283–328.
44. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 2002;**19**(12): 2226–38.
45. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 2001;**55**:709–42.
46. Hollich V, Storm CE, Sonnhammer EL. OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics* 2002;**18**(9):1272–3.
47. Storm CE, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002;**18**(1):92–9.
48. Storm CE, Sonnhammer EL. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* 2003;**13**(10):2353–62.
49. Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002;**3**:14.
50. Dufayard JF, Duret L, Penel S, *et al*. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 2005;**21**(11):2596–603.
51. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**(5338):631–7.
52. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
53. Li L, Stoeckert CJJr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**(9):2178–89.
54. Jothi R, Zotenko E, Tasneem A, *et al*. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* 2006;**22**(7): 779–88.
55. van der Heijden RT, Snel B, van Noort V, *et al*. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 2007;**8**:83.
56. Ruan J, Li H, Chen Z, *et al*. TreeFam: 2008 update. *Nucleic Acids Res* 2008;**36**(Database issue):D735–40.
57. Conte MG, Gaillard S, Droc G, *et al*. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics* 2008;**9**:183.
58. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, *et al*. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 2011; **39**(Database issue):D556–60.
59. Datta RS, Meacham C, Samad B, *et al*. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res* 2009;**37**(Web Server issue):W84–9.
60. Huerta-Cepas J, Dopazo H, Dopazo J, *et al*. The human phylome. *Genome Biol* 2007;**8**(6):R109.
61. Pryszcz LP, Huerta-Cepas J, Gabaldon T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res* 2011;**39**(5):e32.
62. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 2005;**15**(10): 1456–61.
63. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* 1996;**93**(19):10268–73.
64. Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999;**96**(6):2896–901.
65. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
66. Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**: 421.
67. Lopez R, Silventoinen V, Robinson S, *et al*. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res* 2003;**31**(13):3795–8.
68. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**(1):195–7.
69. Wall DP, Fraser HB, Hirsh AE. Detecting putative orthologs. *Bioinformatics* 2003;**19**(13):1710–1.
70. Kuzniar A, van Ham RC, Pongor S, *et al*. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* 2008;**24**(11):539–51.
71. Wolf YI, Novichkov PS, Karev GP, *et al*. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 2009;**106**(18):7273–80.
72. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* 2008;**36**(21):6688–719.

73. King N. The unicellular ancestry of animal development. *Dev Cell* 2004;**7**(3):313–25.
74. Ekman D, Bjorklund AK, Frey-Skott J, *et al*. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* 2005;**348**(1):231–43.
75. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 2001;**310**(2):311–25.
76. Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* 2001;**52**(6):540–2.
77. Tatusov RL, Natale DA, Garkavtsev IV, *et al*. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 2001; **29**(1):22–8.
78. Jensen LJ, Julien P, Kuhn M, *et al*. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 2008;**36**(Database issue):D250–4.
79. Muller J, Szklarczyk D, Julien P, *et al*. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* 2010;**38**(Database issue): D190–5.
80. Kristensen DM, Kannan L, Coleman MK, *et al*. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 2010;**26**(12):1481–7.
81. Mulkidjanian AY, Koonin EV, Makarova KS, *et al*. The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci USA* 2006;**103**(35):13126–31.
82. Makarova K, Slesarev A, Wolf Y, *et al*. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci USA* 2006;**103**(42):15611–6.
83. Makarova KS, Sorokin AV, Novichkov PS, *et al*. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct* 2007; **2**:33.
84. Kristensen DM, Cai X, Mushegian A. Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J Bacteriol* 2011;**193**(8):1806–14.
85. Yutin N, Wolf YI, Raoult D, *et al*. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* 2009;**6**:223.
86. Montague MG, Hutchison CA 3rd. Gene content phylogeny of herpesviruses. *Proc Natl Acad Sci USA* 2000; **97**(10):5334–9.
87. Van Dongen S. *Graph Clustering by Flow Simulation*. The Netherlands: University of Utrecht, 2000.
88. Chen F, Mackey AJ, Stoeckert CJ Jr, *et al*. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006;**34**(Database issue): D363–8.
89. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001;**314**(5):1041–52.
90. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 2005;**33**(Database issue):D476–80.
91. Alexeyenko A, Tamas I, Liu G, *et al*. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 2006;**22**(14):e9–15.
92. Ostlund G, Schmitt T, Forslund K, *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 2010;**38**(Database issue):D196–203.
93. Dessimoz C, Cannarozzi G, Gil M, *et al*. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: McLysath A, Huson DH, (eds). *RECOMB 2005 Workshop on Comparative Genomics*. Dublin, Ireland, *Vol. LNBI 3678 of Lecture Notes in Bioinformatics*, Berlin/Heidelberg, Germany: Springer, 2005; 61–77.
94. Roth AC, Gonnet GH, Dessimoz C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 2008;**9**: 518.
95. Altenhoff AM, Schneider A, Gonnet GH, *et al*. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* 2011;**39**(Database issue): D289–94.
96. Deluca TF, Wu IH, Pu J, *et al*. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* 2006;**22**(16):2044–6.
97. Chen TW, Wu TH, Ng WV, *et al*. DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection. *BMC Bioinformatics* 2010;**11(Suppl 7)**:S6.
98. Linard B, Thompson JD, Poch O, *et al*. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* 2011;**12**:11.
99. Consortium, I.C.G.S Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004;**432**(7018): 695–716.
100. Koonin EV. Evolution of genome architecture. *Int J Biochem Cell Biol* 2009;**41**(2):298–306.
101. Koonin EV, Wolf YI. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* 2010; **11**(7):487–98.
102. Bourque G, Zdobnov EM, Bork P, *et al*. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res* 2005;**15**(1):98–110.
103. Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. *BMC Genomics* 2009; **10**:630.
104. Zdobnov EM, Bork P. Quantification of insect genome divergence. *Trends Genet* 2007;**23**(1):16–20.
105. Zdobnov EM, von Mering C, Letunic I, *et al*. Consistency of genome-based methods in measuring Metazoan evolution. *FEBS Lett* 2005;**579**(15):3355–61.
106. Novichkov PS, Wolf YI, Dubchak I, *et al*. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J Bacteriol* 2009; **191**(1):65–73.
107. Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet* 2001;**17**(1):10–3.
108. Wolf YI, Rogozin IB, Kondrashov AS, *et al*. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 2001;**11**(3):356–72.
109. Huynen MA, Bork P. Measuring genome evolution. *Proc Natl Acad Sci USA* 1998;**95**(11):5849–56.

110. Rogozin IB, Makarova KS, Murvai J, *et al*. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* 2002;**30**(10):2212–23.

111. Rogozin IB, Makarova KS, Natale DA, *et al*. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* 2002;**30**(19):4264–71.

112. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 2003; **1**(2):127–36.

113. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. *Trends Genet* 1996;**12**(9):334–6.

114. Galperin MY, Koonin EV. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica* 1999;**106**(1–2): 159–70.

115. Rolland T, Neuveglise C, Sacerdot C, *et al*. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One* 2009;**4**(8):e6515.

116. Cannon SB, Young ND. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics* 2003;**4**:35.

117. Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* 2006;**2**(9):e133.

118. Wapinski I, Pfeffer A, Friedman N, *et al*. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 2007;**23**(13):i549–58.

119. Novichkov PS, Ratnere I, Wolf YI, *et al*. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res* 2009;**37**(Database issue): D448–54.

120. Fu Z, Jiang T. Clustering of main orthologs for multiple genomes. *J Bioinform Comput Biol* 2008;**6**(3):573–84.

121. Shi G, Zhang L, Jiang T. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinformatics* 2010;**11**:10.

122. Fulton DL, Li YY, Laird MR, *et al*. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 2006;**7**:270.

123. Vashist A, Kulikowski CA, Muchnik I. Ortholog clustering on a multipartite graph. *IEEE/ACM Trans Comput Biol Bioinform* 2007;**4**(1):17–27.

124. Towfic F, VanderPlas S, Oliver CA, *et al*. Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinformatics* 2010;**11(Suppl 3)**:S7.

125. Gabaldon T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 2008;**9**(10):235.

126. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 2009;**5**(1):e1000262.

127. Chen F, Mackey AJ, Vermunt JK, *et al*. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2007;**2**(4):e383.

128. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res* 1998;**7**(4):354–70.

129. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004;**60**(2):427–35.

130. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 2004;**428**(6983):617–24.

131. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 2005;**3**(10):e314.

132. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–9.

133. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;**28**(1):304–5.

134. Liao BY, Zhang J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* 2006;**23**(3):530–40.